Published in final edited form as:

Am J Rhinol. 2008 ; 22(1): 47–51. doi:10.2500/ajr.2008.22.3129.

Anatomical reconstruction from endoscopic images: Toward quantitative endoscopy

Hanzi Wang, Ph.D.^{*}, Daniel Mirota, B.S.^{*}, Gregory Hager, Ph.D.^{*}, and Masaru Ishii, M.D., Ph.D.[#]

^{*} Center for Computer-Integrated Surgical Systems and Technology, The Johns Hopkins University, Baltimore, Maryland [#] Department of Otolaryngology–Head and Neck Surgery, The Johns Hopkins Bayview Medical Center, Baltimore, Maryland

Abstract

Background—Recent advances in computational image processing have made it possible to reconstruct camera motion and scene geometry from a series of monocular images. By applying these methods to endoscopic image sequences, it is possible to create detailed, quantitative anatomic reconstructions. Such anatomic reconstructions have many potential clinical uses. Our objectives in this study are to (1) develop a process flow for reconstruction from endoscopic image sequences and (2) present results supporting the hypothesis that such reconstructions can be computed.

Methods—We first outline the overall process flow for endoscopic reconstruction. Then, we present an instantiation of this process flow using recently developed methods in computational vision. We apply these methods to cadaverous specimens for which ground truth endoscopic motion is known.

Results—We are able to produce consistent estimates of endoscopic motion and dense reconstructions of the surrounding anatomy for >65% of 1373 image pairs.

Conclusion—Our study indicates that processing endoscopic images to produce anatomic structure is feasible. Such reconstructions have high potential clinical value for intraoperative navigation, diagnosis, and treatment planning.

Keywords

Anatomic reconstruction; computer-integrated surgery; computer vision; endoscopy; image processing; navigation; robust statistics; surgery; visual modeling

Computer-assisted surgery is a maturing field that has expanded our ability to perform complex sinus and anterior skull-based surgeries in a safe fashion. However, these new techniques demand new levels of precision in surgical navigation and challenging problems in the visualization of anatomy and anatomic revisions or reconstructions.

In this study, we present a new concept in computer-assisted surgery, which we refer to as “quantitative endoscopy” (QE). QE allows for the precise and quantitative reconstruction of surface anatomy from a series of monocular endoscopic images. QE has been made possible by recent advances in image processing algorithms that reconstruct camera motion and scene

Address correspondence and reprint requests to Gregory Hager, Ph.D., Department of Computer Science, The Johns Hopkins University, Baltimore, MD 21218 hager@cs.jhu.edu.

Presented at the meeting of the American Rhinologic Society, Washington, D.C., September 15, 2007

The authors have no financial interest or other conflict of interest in this work

geometry from a series of monocular images. Solutions to this problem, referred to as structure from motion in the computer vision literature,¹ have found numerous applications in digital imaging. Our goal is to apply these methods to endoscopic image sequences to create detailed, quantitative anatomic reconstructions.

Several groups have specifically used endoscopic video to extract local geometry. Taylor *et al.*^{2,3} used simple hierarchical correlation methods to track identified landmarks in successive images from a robotically manipulated laparoscope. Mori⁴ has developed a virtual endoscopy system making use of epipolar geometry. Several groups, including ours^{5,6} and Mourgues *et al.*⁷ have produced dense depth/surface maps from carefully calibrated stereo endoscopes. The work of Higgins extensively considered the development of CT to video registration for the purpose of lung cancer assessment and virtual bronchoscopy.^{8–11} However, all of this work assumes the presence of a 3D model (from CT), and the goal is to register against this model. Consequently, the methods they have developed are not able to perform explicit surface reconstruction from video data. Our group developed an initial demonstration of motion-based anatomic reconstruction on phantom data.¹²

The development of QE, particularly of tubular structures such as the sinuses, ear canals, or throat, presents a number of technical challenges. Extreme changes in lighting, low image texture, and unfavorable motion trajectories all contribute to the difficulty of the reconstruction problem. Thus, our goals in this study are to (1) provide an instantiation of the feature matching and reconstruction framework suitable for typical sinus imagery, (2) validate the hypothesis that motion estimation can be performed reliably on endoscopic data with this framework, and (3) provide illustrative data showing the resulting anatomic surface reconstructions.

TECHNICAL BACKGROUND

Consider a camera observing a point $P = (x, y, z)^t \in \mathbb{R}(3)$ on a surface S from a position that can be characterized by the location of the optical center c , and a rotation matrix $\mathbf{R}$, expressed relative to a fixed coordinate system. The projection of P onto the camera image plane can be described in two steps. First, the point P is projected into camera coordinates as $P_c(x_c, y_c, z_c)^t = R^t(P - c)$, where R^t denotes the transpose of R . Once in camera coordinates, the point is then projected onto the image plane as $p = (u, v)^t = (fx_c/z_c, fy_c/z_c)^t$, where f is the focal length of the imaging system. In practice, there are additional parameters that identify the location of the optical center in the pixel grid and scale coordinates to pixel coordinates. The identification of these parameters is well studied in the computer vision and photogrammetry literature.¹³

Consider now observing the same physical point P from two camera positions (c_1, R_1) and (c_2, R_2) with $c_1 \neq c_2$. The point P will project to two image locations $p_1 = (u_1, v_1)^t$ and $p_2 = (u_2, v_2)^t$. The following condition now holds¹:

$$(u_2, v_2, 1)^t \text{sk}(c_2 - c_1) R_2^t R_1 (u_1, v_1, 1) = 0 \quad (1)$$

where sk denotes a skew symmetric matrix.¹⁴ The matrix $\mathbf{E} = \text{sk}(c_2 - c_1) R_2^t R_1$ therefore encodes the motion of the camera between the two viewpoints. It is referred to as the *essential matrix*.¹⁵ The essential matrix can be estimated using linear techniques from eight or more matched points in two images¹⁶ or nonlinear techniques from as few as five points.¹⁷ The process of computing $\mathbf{E}$ (or other equivalent forms) and using it to estimate the geometric relationship between the cameras is referred to as the motion problem. It is important to note that if some translation component $c_2 - c_1$ satisfies Eq. 1, then any scalar multiple also does—multiplying

the zero vector on the right side by a scalar has no effect. Thus, translation can be estimated only up to a scale factor using this technique.¹

Given two known camera positions, it also is possible to triangulate observed points and compute their position in 3D space. This is referred to as the stereo problem.¹⁸ The essential issue in stereo is to match corresponding points in the two images. Extensive studies of local matching solutions and globally optimal matching solutions¹⁹ have been published, although no definitive answers exist.¹⁸ However, in most cases it is possible to implement efficient and robust local matching solutions.

If it is the case that both motion and observed scene geometry is unknown and must be jointly estimated, the problem is referred to as the structure and motion problem.²⁰ For the case of two frames, this problem can be solved as follows: (1) find matching points in two images, (2) compute the essential matrix from the matched points, (3) extract the motion (rotation $R_2^1 R_1$ and translation $c_2 - c_1$) between the two positions from the essential matrix; and (3) perform stereo matching and reconstruction to compute the geometry of the observed scene. If an extended series of frames is available (*e.g.*, by tracking points through a video sequence), methods for reconstruction from multiple frames have been developed for both simplified camera models²¹ or full perspective models.²² In particular, bundle-adjustment methods²³ can be applied to extract surface geometry and motion information with extremely high accuracy. This geometry can then be used for a wide variety of clinical applications. The overall process flow for QE reconstruction thus can be viewed as shown in Fig. 1.

METHODS AND MATERIALS

Data Collection

We performed a series of video-assisted nasal endoscopies on a cadaverous porcine specimen. Images were collected using a Storz Telecam 202212113U NTSC, attached to a 4-mm Storz zero-degree rigid Hopkins rod endoscope 7210AA (Karl Storz, Inc., Culver City, CA). The images were digitally captured at a rate of roughly 30 frames/s and stored for later data processing. An Optotrak (Northern Digital, Waterloo, ON, Canada) was used to measure and record endoscopic motion data during image acquisition *via* an Optotrak optical target attached to the shaft of the endoscope. The Optotrak motion data are considered the ground truth to which the reconstructed endoscopic motion data were compared. Optotrak motion data were recorded with timing information to allow for proper synchronization with the image data.

Images from a standard optical calibration target were also recorded before the data collection. An optical calibration procedure¹³ was performed on these images using a publically available Matlab toolkit (Mathworks, Inc., Natick, MA).²⁴ The resulting calibration had an RMS error of 0.33 pixels indicating that an accurate calibration was achieved.

The principal goal of this study was to compare the accuracy of image-based motion estimation with the Optotrak system. To perform this comparison, it is necessary to relate the pose (position and orientation) of the Optotrak target to that of the image sensor. Computing this transformation involves solving a so-called “ $AX = XB$ ” problem. We implemented the method of Park and Martin²⁵ and applied it to the Optotrak pose data acquired during the optical calibration procedure. The resulting $AX = XB$ solution had an RMS error 0.94° of orientation and 0.53 mm of translation error.

Computing Epipolar Geometry from Images

We have developed a reliable method for estimating endoscope epipolar geometry by using scale-invariant feature transform (SIFT) features²⁶ and robust estimation techniques.²⁷ The algorithm proceeds as follows. First, SIFT features are detected and matched in a given pair of images using methods reported by Lowe.²⁶ Five feature matches are then sampled randomly and without replacement, and a candidate **E** matrix is computed using the five-point algorithm.¹⁷ The resulting **E** matrix is normalized and the left side of Eq. 1 is evaluated on all matched pairs. The solution is given a score that is the median of these evaluations. The final solution is the **E** matrix with the least median score. The advantage of this algorithm is that it uses the minimum number of correct matches⁵ and can tolerate up to 50% outliers in the data.

Multiframe Tracking and Reconstruction

Matched features between pairs of images can be used for 3D reconstruction. However, it is well known that more accurate reconstructions result by combining information across multiple images.^{23,12} We have implemented an initial version of such an algorithm.

In this algorithm, SIFT features are extracted and matched as described previously. However, as the sequence proceeds, SIFT features that match across multiple frames are tracked throughout the video sequence. Reconstruction from the tracked features proceeds as follows.

First, two frames are selected, the camera motion is estimated as described previously and the validated matches (*i.e.*, inliers) are obtained. An initial 3D reconstruction (up to an unknown scale) of the matched points is determined through triangulation using the estimated camera motion and the validated matches. For a new frame, matches corresponding to the tracked features are used to infer correspondences between 2D image features and their 3D reconstructed points (known from previous frames). The camera motion is estimated using these 2D–3D matches using standard methods.²⁸ The 3D positions of the reconstructed points are then refined using a least squares technique applied to all frames. Newly reconstructed 3D points corresponding to new SIFT features are initialized and are added to the reconstruction. In this way, a sequence of motions and an associated extended reconstruction is computed.

RESULTS

Figure 2 illustrates the performance of the robust motion estimation algorithm. In the upper row, the initial feature matches from a pair of sinus images are shown. At the lower left, the feature matches that are judged to be inliers after the robust **E** matrix estimation are shown. For each inlier match in the left image, it is possible to compute the line along which the corresponding match in the right image must lie. The intersection of those lines shows the direction of motion, as depicted in the lower right image.

Figure 3 illustrates the steps of the two-frame structure reconstruction algorithm on a typical image pair. On the upper left is the image pair. On the lower right are the epipolar lines indicating, in this case, a lateral motion. The lower middle image shows the results of a dense stereo match, in which brightness encodes depth (brighter is closer). The right-most image shows a view of the 3D reconstruction results, clearly showing the curving tubular shape of the viewed cavity.

To further validate the accuracy of the image matching component, we have applied the algorithm to 1373 images from the porcine video sequence. Each image was paired with a second image at least 1 mm distant from the sequence, and the two-frame algorithm was applied. The interframe motion values from image-based motion estimation and tracked motion were compared. Because the absolute value of translation can not be computed from images alone, the direction of translation was used as the basis of comparison.

It was found that accurate motion information was acquired in frames for which at least 50 feature matches were present. This comprised 855 (65.8%) of the image pairs. For these pairs, the RMS error in rotation was 1.14° with a median error of 0.42° . The error in translation direction was 26.4° and the median error was 14.0° . However, these results were quite variable. For example, the comparable figures for frames 250–350 were 0.47 and 0.26° of rotation and 8.79 and 6.14° of translation. The latter are well within the expected noise floor of motion estimation using the five-point method,¹⁷ and are very reasonable candidates for further refinement.

We have also applied the multiframe tracking and reconstruction algorithm to the data. At this point, we have not performed a quantitative comparison of this algorithm. However, qualitatively, the algorithm results in a larger area of reconstructed points, and the reconstructed data have higher accuracy. Figure 4 shows one such reconstruction result.

DISCUSSION

Our results indicate that traditional endoscopic images can be computer enhanced to produce accurate geometric models of viewed structures. The proposed algorithm was able to generate reliable motion estimates on 65.8% of 1373 contiguous frame pairs and highly accurate motion estimates on 100% of a selected series of 100 frame pairs. Inspection of the resulting reconstructed points showed that the reconstructions are dense and are qualitatively correct relative to the viewed anatomy. We are now validating the accuracy of the geometric reconstruction. This will be performed by registering the reconstructed data to surfaces extracted from a CT image of the anatomy. This process also will provide a demonstration of anatomy-based navigation using video-CT registration.

The technique presented here has been optimized for tubular geometries that severely constrain camera motions, such as those typically found in the head and neck region. However, they are fully generalizable to unconstrained endoscopic procedures. As such, these methods are immediately applicable to the problems of registration for surgical navigation, sizing of anatomic structures, and 3D visualization of surgical anatomy for surgical planning or simulation. QE may also be useful for surgical navigation and updating anatomic models during surgery.

We foresee the integration of QE into current clinical procedures to be simple and straightforward. However, it is clear from our results that performance can be enhanced by careful and appropriate image acquisition. The development of appropriate clinical protocols as well as further technical improvements to the algorithms are the next steps toward clinical application.

Acknowledgments

Supported by Grant NIH 1R21EB005201-01A1

References

1. Longuet-Higgins HC. A computer algorithm for reconstructing a scene from two projections. *Nature* 1981;293:133–135.
2. Taylor RH, Funda J, Eldridge B, et al. Telerobotic assistant for laparoscopic surgery. *IEEE Eng Med Biol* 1995;14:279–288.
3. Taylor, RH.; Funda, J.; Eldridge, B., et al. A telerobotic assistant for laparoscopic surgery. In: Taylor, R.; Lavallee, S.; Burdea, G.; Moesges, R., editors. *Computer-Integrated Surgery*. Cambridge, MA: MIT Press; 1996. p. 581-592.

4. Mori, K.; Deguchi, D.; Hasegawa, J., et al. Proc Int Conf Medical Image Computing and Computer-Assisted Interventions. New York: Springer Verlag; 2001. A method for tracking the camera motion of real endoscope by epipolar geometry analysis and virtual endoscopy system.
5. Burschka D, Corso JJ, Dewan M, et al. Navigating inner space: 3-D assistance for minimally invasive surgery. Robot Atonom Syst 2005;52:5–26.
6. Lau, WW.; Ramey, NA.; Corso, JJ., et al. Proc Int Conf Medical Image Computing and Computer-Assisted Interventions. New York: Springer Verlag; 2004. Stereo-based endoscopic tracking of cardiac surface deformation.
7. Mourgues, F.; Coste-Maniere, E. Proc Int Conf Medical Image Computing and Computer-Assisted Interventions. New York: Springer Verlag; 2002. Flexible calibrations of actuated stereoscopic endoscope for overlay in robot assisted surgery.
8. Helferty JP, Higgins WE. Technique for registering 3d virtual CT images to endoscopic video. ICIP 2001;2:893–896.
9. Helferty JP, Higgins WE. Combined endoscopic video tracking and virtual 3d ct registration for surgical guidance. ICIP 2002;2:961–964.
10. Kiraly AP, Helferty JP, Hoffman EA, et al. Three-dimensional path planning for virtual bronchoscopy. IEEE Trans Med Imag 2004;23:1365–1379.
11. Rai, L.; Higgins, WE. Image-based rendering method for mapping endoscopic video onto CT-based endoluminal views. In: Cleary, KR.; Galloway, RL., Jr, editors. Medical Imaging 2006: Visualization, Image-Guided Procedures, and Display; Proceedings of the Conference of the Society of Photo-Optical Instrumentation Engineers; SPIE; 2006. p. 1-12.
12. Burschka D, Li M, Taylor R, et al. Scale-invariant registration of monocular endoscopic images to CT-scans for sinus surgery. Med Image Anal 2005;9:413–439. [PubMed: 16009593]
13. Zhang Z. A flexible new technique for camera calibration. IEEE Trans Pattern Anal Mach Intell 2000;22:1330–1334.
14. Strang, G. Linear Algebra and Its Applications. 2. New York: Academic Press, Inc; 1980.
15. Ma, Y.; Soatto, S.; Kosecka, J.; Sastry, S. An Invitation to 3-D Vision. New York: Springer Verlag; 2004.
16. Hartley RI. In defense of the eight-point algorithm. IEEE Trans Pattern Anal Machine Intell 1997;19:580–593.
17. Nister D. An efficient solution to the five-point relative pose problem. IEEE Trans Pattern Anal Machine Intell 2004;26:756–777.
18. Brown MZ, Burschka D, Hager GD. Advances in computational stereo. IEEE Trans Pattern Anal Machine Intell 2003;25:993–1008.
19. Scharstein D, Szeliski R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int J Comput Vision 2002;47:7–42.
20. Huang T, Netravali A. Motion and structure from feature correspondences: A review. Proc IEEE 1994;82:252–268.
21. Morita T, Kanade T. A sequential factorization method for recovering shape and motion from image streams. IEEE Trans Pattern Anal Machine Intell 1997;19:858–867.
22. Pollefeys M, Van Gool LJ, Vergauwen M, et al. Visual modeling with a hand-held camera. Int J Comput Vision 2004;59:207–232.
23. Triggs, B.; McLauchlan, P.; Hartley, RI.; Fitzgibbon, AW. Bundle adjustment: A modern synthesis. Lecture Notes in Computer Science; Proc Int Workshop on Vision Algorithms: Theory and Practice; New York: Springer Verlag; 1999. p. 298-372.
24. Bouget, J-Y. The matlab camera calibration toolkit. [last accessed Jan. 3, 2008]. Available online at www.vision.caltech.edu/bouguetj/calib_doc/
25. Park FC, Martin BJ. Robot sensor calibration: Solving $AX = XB$ on the euclidean group. IEEE Trans Robot Automat 1994;10:717–721.
26. Lowe D. Distinctive image features from scale-invariant key-points. Int J Comput Vision 2004;60:91–110.
27. Rousseauw, PJ.; Leroy, A. Robust Regression and Outlier Detection. Hoboken, NJ: John Wiley and Sons; 1987.

28. Hager G, Lu C-P, Mjolsness E. Object pose from video images. *IEEE Trans Pattern Anal Machine Intell* 2000;22:610–622.

Figure 1.
The data flow for systems performing reconstruction from endoscopic digital video.

Figure 2.

A demonstration of the matching and motion estimation algorithm on a pair of images. The upper row shows the image pair and the initial SIFT feature matches. The lower left figure shows the computed inlier matches. The lower right figure shows the motion direction as the intersection of the epipolar lines resulting from the matched features.

Figure 3.

An illustration of dense reconstruction from two images. The image pair and validated matched features are shown on the upper left. Below, the resulting epipolar geometry showing this to be a side-to-side motion. Middle, bottom, the depth of points in the image encoded by brightness (closer is brighter). A 3D reconstruction from the matched image points is shown on the right.

Figure 4.

The results of surface reconstruction after tracking feature points through several frames of video. The first and last frame are shown at the top. The feature tracks are shown overlaid on the final image at the bottom left, and the resulting reconstruction is shown at the bottom right.